International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | ICAIDSCT26-Conference Proceeding
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.ICAIDSCT26-145

Predicting Very High-Cost Claimants Using Symmetry ETG/PEG Feature Engineering Combined with Advanced Machine Learning

Mani Kanta Pothuri
Independent Researcher USA.

Abstract: Symmetry ETP/PEG are also utilized for engineering to examine provider effectiveness, measure care quality, and examine treatment protocols. It is one of the most significant and tool that has been utilized for the purpose of grouping healthcare claims data into essential, patient centric episode of care. Main motive of this systematic review is to determine the key approaches linked with predictive modeling for assessing high-cost claimants in the healthcare industry.

Considering the research methodology, interpretivism research philosophy has taken into account over positivism, realism, and pragmatism. The inductive research approach has also been selected over the deductive approach because it is suitable for conducting a non-statistical investigation. Descriptive research design and grounded theory as research strategy are also chosen to conduct the study in an appropriate and ethical manner. The secondary data collection method has also been utilized to extract information from available sources and databases.

From the findings, it has been determined that healthcare cost concentration is all about the phenomenon where a small extent of the population accounts for an inappropriately large share of overall healthcare spending. There is a great extent of need for episode-based analytics in the healthcare sector, which is understood by considering the growth of the market. The results represented and drawn based on the secondary data have presented and specified the significant of Symmetry ETG/PEG as key episode-based analytics models in the context of predictive healthcare analytics. Organizations should utilize the CatBoost Regressor, SAP Analytics Cloud, and LIME techniques. The primary data collection method can be utilized in the future.

Keywords: ETG, Procedure Episode Groups PEG, High-Cost Claimants, Machine Learning, Healthcare Cost.

1. Introduction

Symmetry ETG/PEG are one of the most significant and useful analytical tools that has been utilized for the purpose of group healthcare claims data into essential, patient centric episode of care. ETG stands for episode treatment groups helps in managing risk and adjusting key activities (Zaleski *et.al* 2025). The best thing about these tools is that they permit analysts to change transactional, raw data into actionable and significant metrics for measuring quality, utilization, and cost. They also contribute to capturing all related services for a particular clinical situation, including inpatient, outpatient, pharmacy and ancillary claims. Symmetry ETP/PEG are also utilized for engineering to examine provider effectiveness, measure care quality, and examine treatment protocols, as well as to predict future healthcare usage. However, the utilization of these tools leads to an increase in the expenditure of organizations established in the healthcare sector for a specific period of time to serve the best care and treatment to patients. The key purpose of this systematic review is to determine the approaches associated with predictive modeling for assessing high-cost claimants in the healthcare industry. At the same time, focus is on determining efficiency of the ETG/PEG feature engineering for boosting provider performance.

2. Methodology

Research methodology refers to the theoretical, significant, and systematic analysis of the principles and methods associated with a branch of knowledge, defining the why behind the collection of data and analysis choices (Dubey & Kothari, 2022). This analysis draws the attention of a person ahead, taking some important and useful methodologies into account. However, this action is based on the type or nature of a study. Thus, the research is qualitative, which means the methods are selected accordingly. For example, the interpretivism research philosophy is taken into account over positivism, realism, and pragmatism. The main reason to select this methodology is to gain in-depth comprehending or knowledge of the selected area of interest.

Along with that, the inductive research approach is also selected over the deductive approach because it is suitable for conducting a non-statistical investigation. The primary motive to choose this approach is that it helps in understanding the relationship between key variables covered in the topic, including symmetry ETG/PEF feature engineering, high-cost claimant prediction, and machine learning. The inductive method is useful as it also contributes to developing new theories based on

specifically observed and detailed information (Cheong *et.al* 2023). The best thing about this approach is that it is open-ended, permitting for the finding of unpredictable insights without being constrained by rigid, existing frameworks or models.

Descriptive research design and grounded theory as research strategy are also chosen to conduct the study in an appropriate and ethical manner, which is quite essential for ethical and timely completion of the overall investigation. The rationale for selecting the descriptive methodology is to collect a detailed, straightforward, and comprehensive summary of the overall situation or phenomena. The research strategy is also selected to build new concepts and models from scratch and represent them as meaningful data.

The secondary data collection method is also utilized to extract information from available sources and databases. It includes academic books, journals, articles, online publications, PubMed, Google Scholar, and Scopus. In addition, to find out the relevant reading materials, key word search tactic is also used. PRISMA is also utilized in the study to improve the completeness, transparency, and consistency of systematic reviews. By using this framework or model, relevant and useful secondary sources have been selected, enabling the collection of a vast amount of data with high accessibility.

Table 1: Inclusion and Exclusion Criteria

Inclusion criteria	Exclusion criteria
Articles after 2020	Articles before 2020
Articles, books and other materials in English language	Articles in other languages
Articles containing keywords like “Symmetry ETG/PEG Feature Engineering, advanced machine learning, and high-cost claimants	Articles which are not containing these keywords

Along with the PRISMA framework, the inclusion and exclusion criteria are also followed and taken into account to improve the quality of data. The reason to use this criterion is to define the topic that has been initiated to study, ensuring the findings are valid, safe, and reliable. They contribute to enhancing the quality of the research by reducing confounding variables, enhancing concentration on the study aim, and maintaining the privacy of other authors and researchers.

3. Results: Evidence Synthesis on Symmetry ETG/PEG Feature Engineering and Machine Learning

3.1. Healthcare cost concentration and the need for episode-based analytics

Healthcare cost concentration is all about the phenomenon where a small extent of the population accounts for an inappropriately large share of overall healthcare spending. Thus, it is determined that health expenditures (HE) are mostly drive by an aging population with those above 65, accounting for 37% of spending despite being a smaller demographic group. Per person personal healthcare investment for the 65 and older people was 22,356 in 2020 and over five times increase than spending per child and almost 2.5 times the spending per working age individual (CMS, 2026).

Table 2: Healthcare cost concentration

Population Segment	Share of Total Healthcare Spending
Top 1%	~24%
Top 5%	~55%
Top 10%	~68%
Bottom 50%	~3%

According to the above table top 1% of users are spending 24% (Mitchell, 2016), the top 5% are spending 55, the top 10 are also conducting the same practice at 68%, and the bottom 50% of patients are spending 3% in healthcare services. There is a high extent of need for episode-based analytics in the healthcare sector, which is understood by considering the growth of the market.

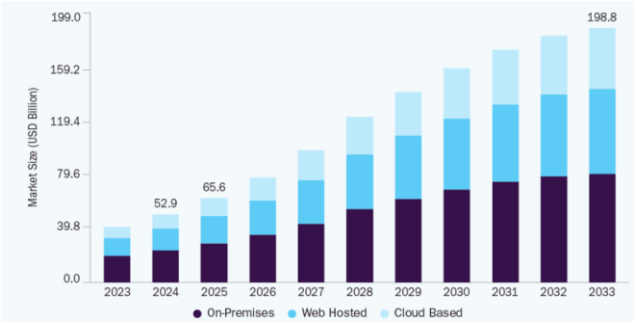

Figure 1: Healthcare Analytics Market

(Source: GVR, 2025)

Thus, as per the above graph, the global healthcare analytics market is expected to grow from 52.98 billion as of 2024 to 198.79 billion by 2033, with a CAGR of 14.85% (GVR, 2025).

3.2. Outline of Symmetry ETG and PEG episode grouping systems

Symmetry ETG and PEG are helpful analytical patient classification software technologies or tools introduced by Optum Business to bundle medicinal rights into defined, longitudinal episodes of care (Elton & Zhang, 2023). They help healthcare firms to analyses expenses, measure performance, and compare quality across the overall care continuum for ETG or PEG. ETG is effective as it allows individuals to emphasize treating an illness appropriately and managing claims into systematic and comprehensive manner. It is a fact that ETG accounts for patients' severity by recognizing comorbidities and problems, permitting for equal and appropriate comparison of provider performance. The purpose of this tool is to manage chronic disease.

Table 3: Comparative table

Feature	ETG	PEG
Focus	Medical condition episodes	Procedure-driven episodes
Data sources	Claims, diagnosis codes	Procedural claims
Purpose	Chronic disease management	Procedure cost analysis
Analytical benefit	Longitudinal treatment tracking	Surgical cost prediction

PEG also contributes to providing and obtaining expected results as they help care providers to track the whole range of care, including pre-operative workups, post-operative follow-up, and the procedure itself (Zaleski *et.al* 2025). Thus, it aids companies to analyses expenditure, clinical performance, and quality, which need to be improved over the last few months or years. The motive of this technology is to conduct the process of cost analysis.

3.3. ETG/PEG feature engineering for predictive modeling

Featuring engineering (FE) is considered the core element of predictive modeling (PM) because it includes the ability to change raw data into important features that shed light on patterns, finally improving the accuracy and interpretability of machine learning (ML) frameworks. It is examined that instead of depending solely on algorithms to determine patterns, FE leverages domain understanding to create, chose and change variables, often making it the main factor in successful forecasting (Sheha *et.al* 2022). ETG/PEG FE permits ML algorithms to learn new things more effective, often resulting in better predictive power than merely tuning algorithm hyper-parameters. Techniques such as domain-specific grouping, like ETG/PEG or polynomial features, also allow models to cover nonlinear connections and unseen patterns that every single raw variable might miss. FE helps in addressing missing values, noise, and outliers, which are common in raw data (Ozdemir, 2022).

Table 4: Examples of ETG/PEG feature variables

Feature Category	Example Variables
Episode frequency	Number of ETG episodes per year
Episode severity	Severity level of treatment episode
Episode duration	Length of treatment episode
Cost intensity	Average cost per episode
Recurrence	Repeat episodes of chronic conditions

It is also effective in connecting prediction tasks or activities for modeling periodic patterns. FE also contributes to increasing model interpretability. As per the above table, episode frequency, severity, duration, cost intensity, and recurrence are noticeable features of the utilized tools.

3.4. Integration of ETG/PEG features with machine learning models

ML is determined as a subset of AI, which aims to develop algorithms to determine types or patterns in data and make forecasts and decisions without being explicitly programmed for every task (Acharya *et.al* 2024). The best thing about this subset is that it enables systems or software to enhance their performance over time by learning from data, experience, and powering advanced technologies. Integration of ETG features with ML models plays a vital role. For example, models permit a person to transform relevant data, transactional claims information into a systematic and structured, and clinically appropriate patient journey.

Table 5: Examples of ML models

Models	Benefits	Application
Random Forest	Covers non-linear interactions	Cost prediction
Gradient Boosting	Appropriate predictive accuracy	Risk stratification

Logistic Regression	Manage complex patterns	Longitudinal modeling
Neural Networks	Interpretable baseline	Benchmark model

The above table defines four different types of ML models, along with their benefits and applications that help in reaching the target and setting new goals.

3.5. ETG/PEG features for identifying very high-cost claimants

ETG also play crucial role in determining very high-cost claimants (HCC) as it is a patient classification system work to determine and examine claims by grouping care services, such as labs, doctor services, prescription drugs, and imaging, into distinct, relevant episodes of care. ETG is highly useful for identifying very HCC by offering a granular, patient-level perspective of spending for more than 500 different situations, permitting insurers to determine high expense outliers (Elton & Zhang, 2023). It is examined that the system or technology creates clinically similar units, which enables more relevant and condition-based cost analysis compared to merely considering the sum annual spending. HCC are those individuals who spend more than the expectation of amount to access healthcare services. The usage of ETG determines these people with significant annual healthcare spending driven by cancer and other chronic diseases.

3.6. Explainable AI for episode-based predictive models

A specialized approach that makes AI systems more effective is called explainable AI for episode-based predictive models (EBPMs). They help in operating as black boxes, trustworthy, transparency and interpretable when forecasting results based on time-bound events, distinct situations, or episodes (Smierzchała *et.al* 2023). The approach is usually important in high-stakes fields like healthcare, where comprehending why a model predicts a specific result that is important for clinical adoption. EBPMs are essential as they help in examining data with a particular temporal window and focus on explaining predictions for single instances. Explainable AI also aids clinicians and financial analysts to comprehend the circumstance and move ahead simple and appropriate way to verify the reasoning (Sheha *et.al* 2022). They also contribute to the following regulations by offering auditable trails of how an episode was examined. Explainable AI prevents individuals from blindly adhering to AI suggestions by visualizing where EBPMs are confident and where they are not.

3.7. Implications for provider performance improvement

The practical implications of predictive models or frameworks in the healthcare context leverage ML and AI, as they enable these technologies to examine historical data, based on which service provider performance can be improved or enhanced (Bolarinwa *et.al* 2025). It enables them to change from reactive to proactive care, thereby enhancing operational effectiveness, patient result and cost management. Key implications include enhanced clinical decision-making process, reduced readmissions, and optimized resource allocation through implementation, which needs careful management of data quality, clinician trust, and ETG/PEG features. By analyzing patient information, models contribute to determining high-risk individuals for targeted discharge planning, which effectively reduces readmission levels (Ozdemir, 2022). Healthcare firms highly use predictive models because it supports them to forecast patient inflow, staffing requirement and bed occupancy.

4. Discussion

The results represented and drawn based on the secondary data have presented and specified the significant of Symmetry ETG/PEG as key episode-based analytics models in the context of predictive healthcare analytics. Thus, it has been determined that the use of these models or tools has transitioned from considering healthcare as disjointed, fee-for-service, unit-based billing to reviewing it as a longitudinal, understandable patient journey. By dividing clinical or medical data, including procedures, pharmacy, and claims, into coherent series, they provide the important granular data to systematically forecast future expenses, patient risks, and resource requirements, along with utilization (Smierzchała *et.al* 2023). Symmetry ETG underpins ERG and SRE, which have permitted firms to examine current risk and predict future challenges, as well as healthcare usage, especially for high-cost, chronic, and complex health issues or complications.

It is also identified that the implementation and utilization of these models is enabled by defining, examining, comparing, and measuring the overall cost and quality of care across the whole continuum, from pre-surgical work-up to post-acute care. The effective utilization of ETG/PEG in the sector has also enabled companies to conduct the benchmarking practice for comparing specialists based on the total expenditures and quality of their processes, including readmissions, complications, and more, instead of just focusing on the cost of medical or treatment procedures. Along with that, it is also analyzed that the use of these tools in clinical practices and services provides specialized analytics, which are far more relevant for predicting expenses than standard, non-clinical claims information. By examining the overall incident, service providers grab the chance to determine and reduce unrequired, high-cost services or unnecessary resources in care that has enabled the venture to obtain the benefit of improving patient outcomes and effectiveness over the last few years (Kumar, 2025).

It is a fact that for any company in the healthcare industry seeking transparency into the quality and cost of healthcare delivery, determining clinical incidents of illness and the services included in their management, diagnosis, and treatment is considered an important business need. Symmetry ETG utilizes routinely gathered claims, including prescriptions, offered

during the course of a patient's treatment, to capture the right services. It also helped in managing the data into meaningful sets of care that result in appropriate identification of clinically homogenous, risk management groups that defines compete care of patients. It has been reflected in the above findings that ETG is useful as it provides a reliable and steady measurement tool for gauging the provision and financing of health care services. It has also enabled firms to execute an array of approaches and initiatives by serving as an analytical unit for examining and comparing the use and financial performance of providers. The tool works by diagnosing codes, including processes and drugs. The system also supports improving effectiveness that helps in providing on time care.

5. Recommendations

Based on the overall findings or results, it is recommended that organizations operating and running in the healthcare sector should utilize the CatBoost Regressor, which is one of the most leading forms of machine learning technologies or methodologies. CatBoost Regressor is noticed as the most advanced ML algorithm usually well-suited for predicting very HCC due to its excellent handling of categorical data, which is prevalent in clinical claims data. This recommended technology can help in building an ensemble of action and decision trees that can enable firms to predict continuous cost amounts. It can also aid in handling categorical features, including procedure codes, ICD-10 diagnosis codes, and provider IDs directly, leading to increase accuracy and less manual FE (Hamid *et.al* 2025). CatBoost Regressor can utilize symmetric trees and ordered boosting that may significantly reduce over-fitting, a basic problem when learning and training on an imbalanced high-cost database. It is also considered a state-of-the-art framework for tabular datasets, rapidly outperforming other frameworks or models in predictive and regression tasks for healthcare expenditures.

It is also suggested that the management should utilize SAP Analytics Cloud to make an appropriate prediction of very high-cost claimants. The suggested tool can play an important role in obtaining expected results as it can provide a robust, low-code source or channel for forecasting HCC by leveraging automated ML though it's smart predict feature. Moreover, it is a fact that SAP Analytics Cloud permits healthcare insurers, payers, and providers to determine members probable to experience important expenses, enabling proactive care management that can help in reducing financial risk (Kumar, 2025). SAP Analytics Cloud is a significant tool for determining HCC as it classifies patients based on their propensity to become high-cost, generating a sorted list of individuals for targeted intervention. It is used by many companies or ventures to predict the real statistical value of future claims or expenditures for a specific group of the population. The tool can support in determining key drivers of expenses, which will be important for firms to determine in a timely manner.

Along with the above recommendations, healthcare firms can also take another suggestion into account. Thus, it is recommended that they should emphasize local, interpretable approximations (LIME) techniques. By using these techniques, they can determine very high-cost health insurance claimants as they translate challenging, black-box ML models into understandable, meaningful, and actionable insights. LIME covers the gap of lack of transparency as they highlight which particular patient features caused a high-cost prediction (Acharya *et.al* 2024). Instead of just flagging a person as high cost, these techniques or methods can explain why. Thus, they can reveal that individuals are flagged due to a blend of musculoskeletal or any other form of disorders and high-cost medicine, permitting for required care management. The utilization of LIME can build trust and create transparency between two or more parties, as in healthcare, comprehending the reasoning or purpose behind a prediction is important for claimers and practitioners to trust.

6. Future research

To conduct the research based on a similar area of interest or topic, the primary data collection method can be utilized in the future, as it can help in addressing the core limitation of the study, which is outdated data. Thus, this methodological term can drive the concentration of a person toward accessing real-time first-hand data collection benefits, which can be tailored to a particular future research aim and objectives. The most considerable fact or benefit about this method is that it can allow capturing immediate patterns or trends and opinions of individuals, rather than relying on stale or historical data. Methods or techniques like online surveys and observational research allow collecting data as events happen, which ensures the in-depth insights reflect the current context more than a past situation. In addition, in the future, the study can be conducted based on a specific nation, as it can help to increase the scope and worth of the research and also contribute to collecting particular statistics, which will define the real situation. A country-specific study can help individuals to consider facts and gain relevant data that can support taking important actions. This way can also reflect the real significance and usefulness of Symmetry ETG/PEG FE.

7. Limitations

Limited control over data quality is the first limitation, as the secondary study depends on existing healthcare claims datasets, due to which a person has no control over how the information is originally collected, analyzed, and interpreted. Mistakes in claims coding and inconsistent reporting may affect the consistency of predictive models. The second limitation is the lack of relevant variables, as the accessibility of datasets may not include all variables required for appropriate prediction. Thus, it may limit the performance of ML models and the efficiency of Symmetry ETG/PEG-based FG. The third limitation is outdated data, as the available sources may contain historical claim information that may not show current healthcare costs,

policy changes, and treatment rules or protocols. As a result, predictive models created from such information may have reduced appropriateness for the current healthcare environment or situation. The worst thing about this limitation is that it influences the quality of data and the process of drawing key findings in the end, which is not suitable for the learning and understanding of readers and others. There are many ways that would be taken into account for the purpose of overcoming the negative impact of the above limitations and eliminating the reasons behind their reemergence.

8. Conclusion

Based on the above analysis, it is concluded that the predictive analysis process has been enhanced or improved by using symmetry ETG/PEG features engineering, combined with advanced ML. It is analyzed that the proper utilization of these tools in the healthcare context has enhanced the effectiveness of decision-making, enabled firms to make the right decision and take the correct action to provide care services to users, as per their expectations. Moreover, by summing up the above investigation, it is summarized that integration of ETG with ML different forms of ML models, including random forest, gradient boosting, and others support increasing the efficiency of operations over the last few weeks or months. It is also examined that the consideration of practical suggestions also aids in improving the process of predicting very HCC that can outcome in terms of speedy recovery of patients and saving their time, along with efforts and money.

References

1. Acharya, N., Kar, P., Ally, M., & Soar, J. (2024). Predicting co-occurring mental health and substance use disorders in women: an automated machine learning approach. *Applied Sciences*, *14*(4), 1630. https://doi.org/10.3390/app14041630
2. Bolarinwa, D., Egemba, M., & Ogundipe, M. (2025). Developing a predictive analytics model for cost-effective healthcare delivery: A conceptual framework for enhancing patient outcomes and reducing operational costs. *International Journal of Advanced Multidisciplinary Research and Studies*, *5*(2), 227-238. https://doi.org/10.62225/2583049X.2025.5.2.3832
3. Cheong, H. I., Lyons, A., Houghton, R., & Majumdar, A. (2023). Secondary qualitative research methodology using online data within the context of social sciences. *International Journal of Qualitative Methods*, *22*, 16094069231180160. https://doi.org/10.1177/16094069231180160
4. CMS, (2026). *NHE Fact Sheet.* [Online]. Retrieved Through: < https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet#:~:text=NHE%20by%20Age%20Group%20and,by%20age%20in%20downloads%20below>. [Retrieved on: 11th March 2026]
5. Dubey, U. K. B., & Kothari, D. P. (2022). *Research methodology: Techniques and trends*. Chapman and Hall/CRC. file:///C:/Users/hp/Downloads/10.1201_9781315167138_previewpdf%20(1).pdf
6. Elton, D., & Zhang, M. (2023). Neck pain service utilization and costs: association with timing of non-pharmaceutical services for individuals initially contacting a primary care provider. A retrospective cohort study. *medRxiv*, 2023-01. https://doi.org/10.1101/2023.01.10.23284193
7. GVR, (2025). *Healthcare Analytics Market*. [Online]. Retrieved Through: https://www.grandviewresearch.com/industry-analysis/healthcare-analytics-market#:~:text=The%20global%20healthcare%20analytics%20market,of%20patient%20retention%20and%20engagement. [Retrieved on: 10th March 2026]
8. Hamid, M., Hajjej, F., Alluhaidan, A. S., & bin Mannie, N. W. (2025). Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling. *Scientific reports*, *15*(1), 31199. https://doi.org/10.1038/s41598-025-13790-x
9. Kumar, R. (2025). Design of a Secure SAP-Enabled Cloud Lakehouse for AI-Driven Financial Risk and Healthcare Analytics. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, *8*(5), 12803-12810. https://doi.org/10.15662/wn7pnz25
10. Mitchell, M, E, (2016). *Concentration of Health Expenditures in the U.S. Civilian Noninstitutionalized Population, 2014*. [Online]. Retrieved Through: < https://meps.ahrq.gov/data_files/publications/st497/stat497.shtml >. [Retrieved on: 11th March 2026]
11. Ozdemir, S. (2022). *Feature engineering bookcamp*. Simon and Schuster. https://books.google.co.in/books?hl=en&lr=&id=xQGEEAAAQBAJ&oi=fnd&pg=PA1&dq=+Episode+Treatment+Groups++feature+engineering+for+predictive+modeling&ots=_jkMjypnEl&sig=i7bPM2wYm3KvR6VX-DGwcCcWgDU&redir_esc=y#v=onepage&q&f=false
12. Sheha, M. A., Mabrouk, M. S., & Sharawy, A. A. (2022). Feature engineering: Toward identification of symptom clusters of mental disorders. *IEEE Access*, *10*, 134136-134156. https://doi.org/0.1109/ACCESS.2022.3232075
13. Smierzchała, Ł., Kozłowski, N., & Unold, O. (2023). Anticipatory classifier system with episode-based experience replay. *IEEE Access*, *11*, 41190-41204. https://doi.org/10.1109/ACCESS.2023.3269879
14. Zaleski, A. L., Guan, X., Thomas Craig, K. J., Junk, C., McGill, A. T., Gordon, H., ... & Caya, K. (2025). An episode-based cost analysis of virtual-first versus in-person-first care to treat common acute conditions among members of a large national payor. *BMC Health Services Research*, *25*(1), 994. https://doi.org/10.1186/s12913-025-13154-1.